

# CORRELAÇÃO LINEAR

TÉCNICAS EM CLIMATOLOGIA  
PROGRAMA DE PÓS-GRADUAÇÃO EM GEOGRAFIA FÍSICA  
3-10 FEV 2017

# Correlação linear

## Referência

Cap. 7 - Métodos Estatísticos para Geografia  
autor: Peter A. Rogerson

- Permite verificar se duas variáveis independentes estão associadas uma com a outra
- Questionamentos iniciais:

“A temperatura de superfície dos oceanos tem alguma relação com a vazão de rios?”

Ou, “a diminuição do preço de um produto tem relação com o aumento de sua oferta? Podem, em um primeiro momento, ser observada através da correlação linear?”

# COEFICIENTE DE CORRELAÇÃO $r$

- Uma das formas utilizadas para se encontrar essas relações é o cálculo do coeficiente de correlação linear de Pearson,  $r$

$$r [-1,0; +1,0]$$

$r = 1,0 \rightarrow$  correlação positiva perfeita

$r = -1,0 \rightarrow$  correlação negativa perfeita

# COEFICIENTE DE CORRELAÇÃO $r$

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

$t_i$	$x_i$	$y_i$
1	$x_1$	$y_1$
2	$x_2$	$y_2$
...	....	....
$t_n$	$x_n$	$y_n$

observações

$$\sum_{i=1}^N$$

→ Somatória

$$x_i \quad y_i$$

→ VETORES ( $x_1, x_2, \dots, x_n$ ) e ( $y_1, y_2, \dots, y_n$ ) - duas variáveis observadas em cada observação, por exemplo, a cada passo de tempo  $i$

$$\bar{x} \quad \bar{y}$$

→ média da amostra  $x$  e de  $y$

$$\sigma_x \quad \sigma_y$$

→ desvio padrão das amostras  $x$  e  $y$

# SOMATÓRIA

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Numerador:

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = (x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y}),$$

$$i = 1, \dots, n$$

**COMO PODE SER ESCRITO O DENOMINADOR?**

# DESVIO PADRÃO $\sigma$ $s$ $dp$

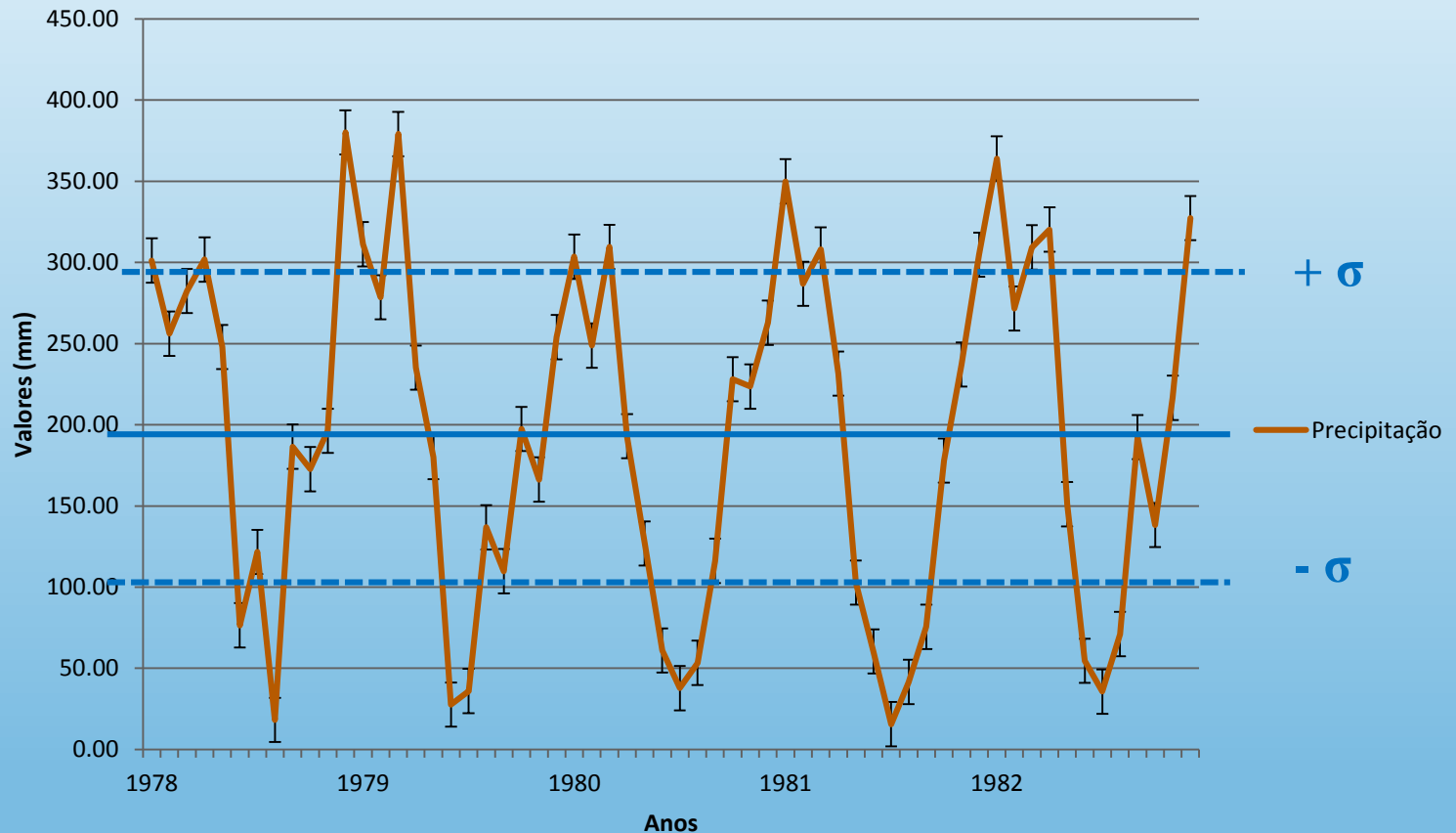
- É uma medida de dispersão e indica a dispersão média de um conjunto de dados em relação à média aritmética da amostra
- Variância = var =  $s^2$   
variância = desvio padrão ao quadrado

# DESVIO PADRÃO

$$dp = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

# Desvio Padrão - exemplo

## Precipitação Mensal

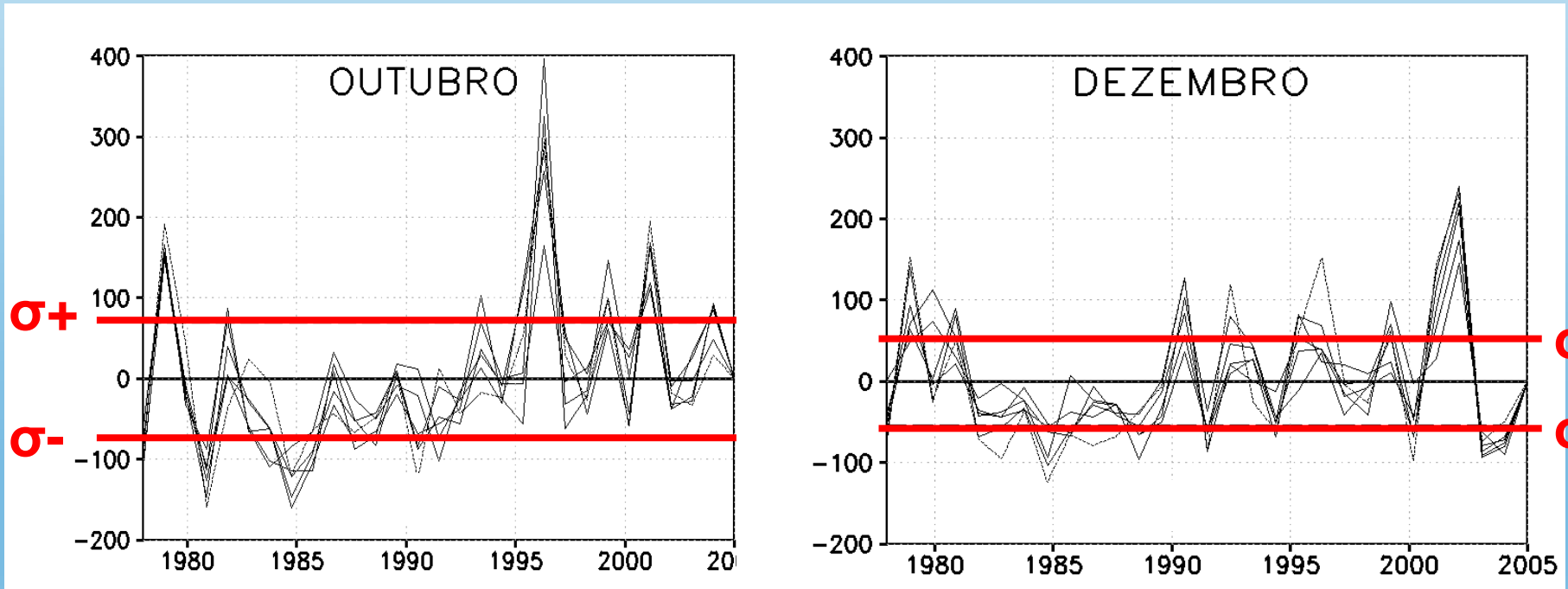


$\sigma = 105,6634$   
pcp média= 194,36  
 $\sigma^2 = 11.164,77$

Dada uma série temporal,  
quantos valores de desvio padrão tem a série?



# ANOMALIA PRECIPITAÇÃO NO NOROESTE DO RS 1978-2005



Sleiman (2005)

# VARIÂNCIA $\sigma^2$

A variância mostra o quão distantes os valores amostrais estão da média, é expressa por:

$$s^2 = \text{var} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

# INTERPRETAÇÃO DA CORRELAÇÃO ENTRE DUAS VARIÁVEIS

- **Correlação positiva**

Quando uma variável aumenta (diminui), a outra também aumenta (diminui)

→ relação diretamente proporcional

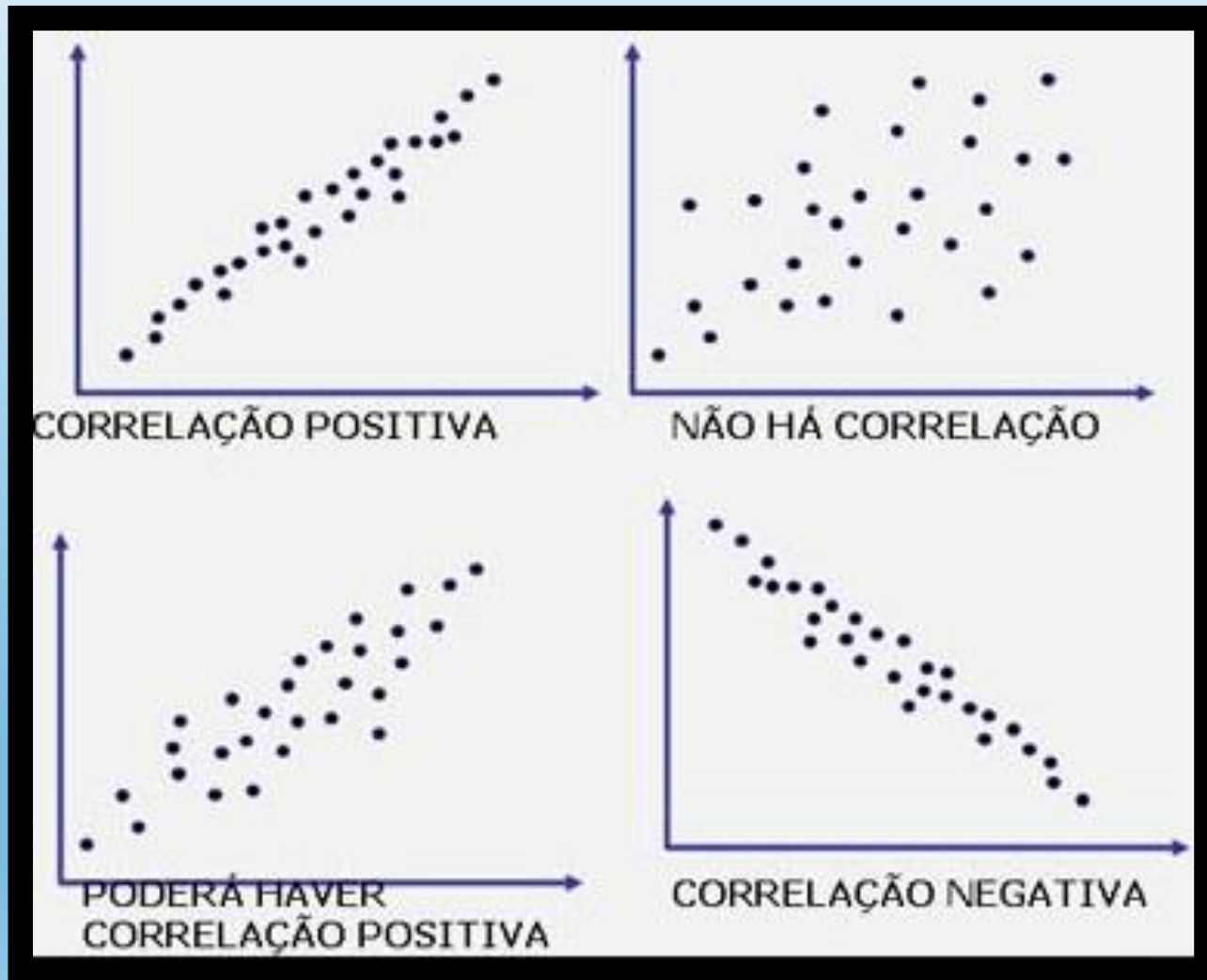
- **Correlação negativa**

Quando uma variável aumenta (diminui), a outra diminui (aumenta)

→ relação inversamente proporcional

- **Sem correlação  $r \rightarrow 0$**

# EXEMPLOS HIPOTÉTICOS DE CORRELAÇÃO ENTRE VARIÁVEIS ALEATÓRIAS



# EXEMPLOS

- Faremos alguns exercícios simples de correlação utilizando uma planilha eletrônica, como Excel ou Calc do BrOffice

Os exemplos dados a seguir foram criados a partir do Excel

# EXERCÍCIO 01: Cálculo da correlação (r) para a amostra de dados de renda e Nível de Educação

- 1) Clique na célula D2
- 2) Na barra de ferramentas, selecione:

Fórmulas – Mais Funções - Estatística - **CORREL**

The screenshot shows the Microsoft Excel interface with the following data in the spreadsheet:

	A	B	C	D	E	F	G	H
		Renda						
1	Observação	(\$x1000)	Educação	r				
2	1	30	12					
3	2	28	12					
4	3	52	18					
5	4	40	16					
6	5	35	16					
7								
8								
9								
10								

The 'Fórmulas' ribbon is active, and the 'Mais Funções' dropdown menu is open, showing the following options:

- Estatística
- Engenharia
- Cubo
- Informações

The 'CORREL' function is highlighted in the list of functions.

## EXERCÍCIO 02

- 1) Clique na célula D2;
- 2) Na barra de ferramentas, selecione:

Fórmulas – Mais Funções - Estatística - **CORREL**

The screenshot shows the Microsoft Excel interface with the 'Fórmulas' ribbon selected. The 'Mais Funções' (More Functions) button is clicked, opening a dropdown menu. The 'Estatística' (Statistics) category is selected, which opens a sub-menu. In this sub-menu, the 'Engenharia' (Engineering) category is selected, which opens another sub-menu. In this final sub-menu, the 'CORREL' function is highlighted. In the background, the spreadsheet shows data for 'Número de corridas vencidas pelo Jôquei principal' (Number of races won by the main jockey) in column D. Cell D2 is highlighted with a red circle and a black border, and a red '1' is next to it. A red '2' is next to the 'CORREL' function in the list.

	A	B	C	D	E	F	G
			Número de corridas vencidas pelo Jôquei principal				
1	Ano	Renda Mediana					
2	1984	35.175	399				
3	1985	35.778	459				
4	1986	37.027	429				
5	1987	37.256	450				
6	1988	37.512	474				
7	1989	37.997	598				
8	1990	37.343	364				
9	1991	36.054	430				
10	1992	35.593	433				
11	1993	35.241	410				
12	1994	35.486	317				
13							

- 3) Na caixa que se abrirá, o campo Matriz1 deverá ser preenchido com os dados referentes à coluna com a renda mediana, ou seja, Coluna B2:B12;
- 4) O mesmo procedimento deverá ser realizado para a Matriz2, porém com os dados do número de corridas, Coluna C2:C12.

dados-curso.xlsx - Microsoft Excel

Início Inserir Layout da Página Fórmulas Dados Revisão Exibição Desenvolvedor

Inserir Função AutoSoma Usadas Recentemente Financeira Lógica Texto Data e Hora Pesquisa e Referência Matemática e Trigonometria Mais Funções Gerenciador de Nomes Definir Nome Usar em Fórmula Criar a partir da Seleção Nomes Definidos Rastrear Preced Rastrear Depen Remover Setas

CORREL  $\text{=CORREL(B2:B12;C2:C12)}$

	A	B	C	D	E	F	G	H	I	J	K	L	M
			Número de corridas vencidas pelo Jôquei principal	r									
1	Ano	Renda Mediana											
2	1984	35.175	399	C2:C12)									
3	1985	35.778	469										
4	1986	37.027	429										
5	1987	37.256	450										
6	1988	37.512	474										
7	1989	37.997	598										
8	1990	37.343	364										
9	1991	36.054	430										
10	1992	35.593	433										
11	1993	35.241	410										
12	1994	35.486	317										
13													
14													
15													
16													

Argumentos da função

CORREL

Matriz1 B2:B12 = {35175;35778;37027;37256;37512;379

Matriz2 C2:C12 = {399;469;429;450;474;598;364;430;43

= 0,558491081

Retorna o coeficiente de correlação entre dois conjuntos de dados.

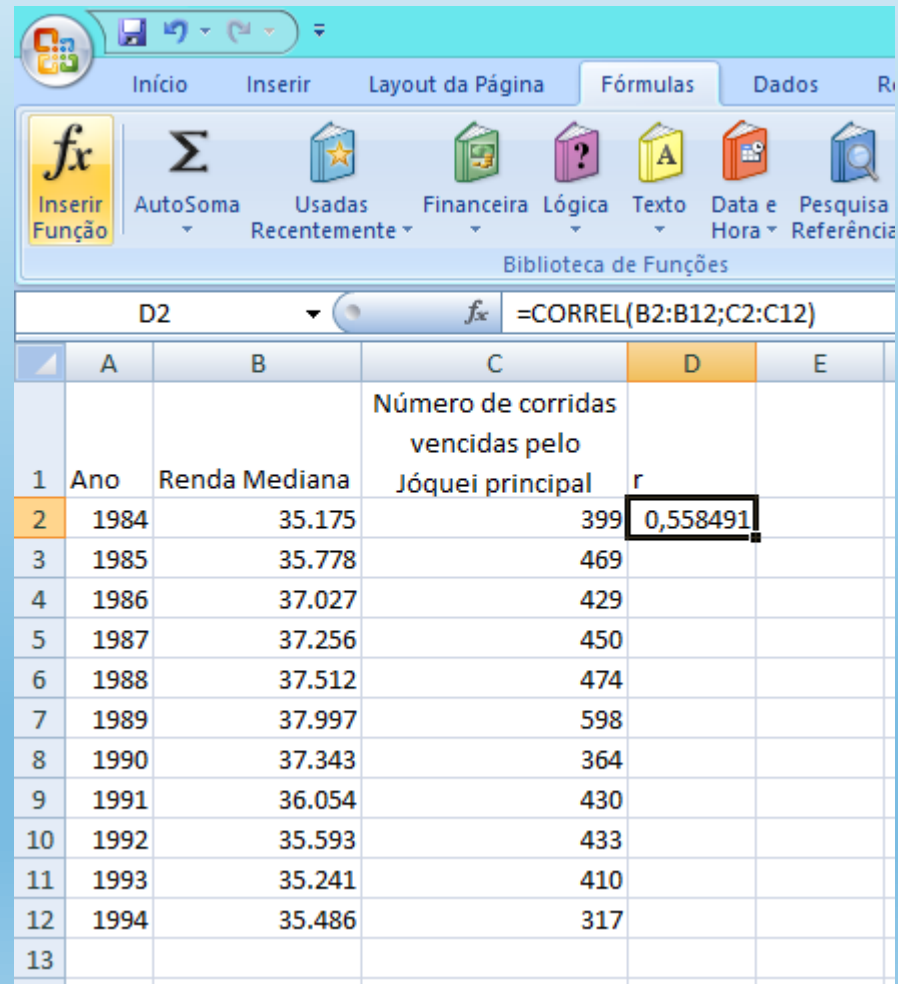
Matriz2 é um segundo intervalo de células de valores. Os valores devem ser números, nomes, matrizes ou referências que contenham números.

Resultado da fórmula = 0,558491081

[Ajuda sobre esta função](#) OK Cancelar



Aperte “OK” para finalizar  
O resultado aparecerá na célula D2

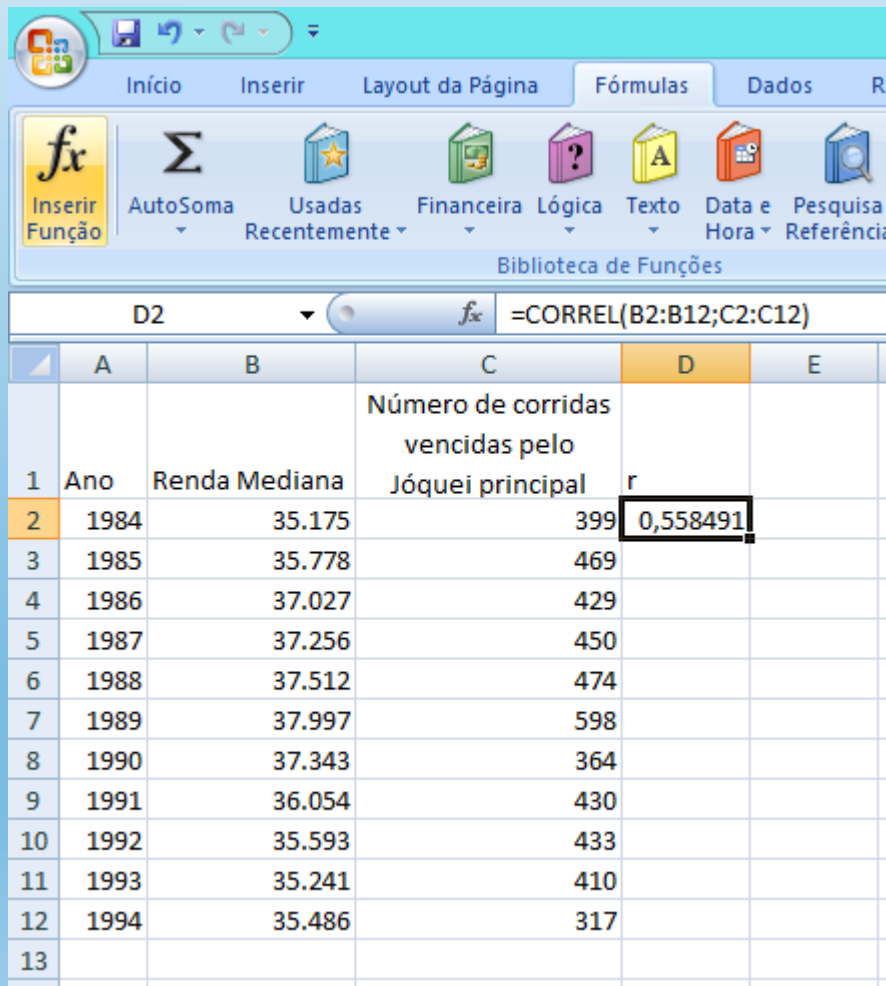


The screenshot shows the Microsoft Excel interface with the 'Fórmulas' ribbon selected. The formula bar displays the formula `=CORREL(B2:B12;C2:C12)`. The spreadsheet data is as follows:

	A	B	C	D	E
			Número de corridas vencidas pelo Jockey principal	r	
1	Ano	Renda Mediana			
2	1984	35.175	399	0,558491	
3	1985	35.778	469		
4	1986	37.027	429		
5	1987	37.256	450		
6	1988	37.512	474		
7	1989	37.997	598		
8	1990	37.343	364		
9	1991	36.054	430		
10	1992	35.593	433		
11	1993	35.241	410		
12	1994	35.486	317		
13					

# INTERPRETAÇÃO DO VALOR GERADO

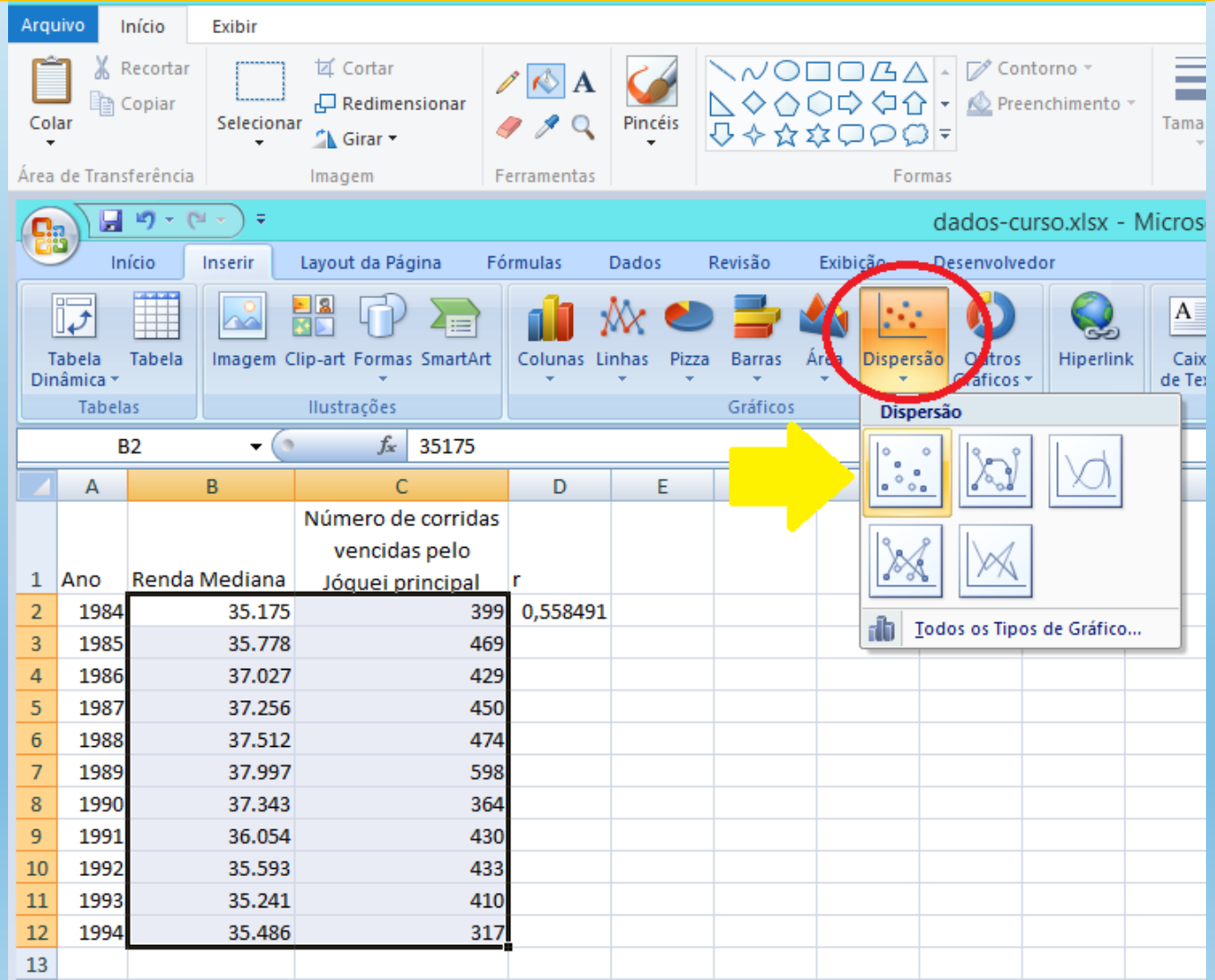
Para a série aleatória gerada nos exemplos, o valor de correlação retornado foi 0,558491



	A	B	C	D	E
			Número de corridas vencidas pelo Jôquei principal	r	
1	Ano	Renda Mediana			
2	1984	35.175	399	0,558491	
3	1985	35.778	469		
4	1986	37.027	429		
5	1987	37.256	450		
6	1988	37.512	474		
7	1989	37.997	598		
8	1990	37.343	364		
9	1991	36.054	430		
10	1992	35.593	433		
11	1993	35.241	410		
12	1994	35.486	317		
13					

Se retornarmos à explicação anterior sobre o coeficiente de correlação, verificamos que as séries possuem alguma correlação linear positiva.

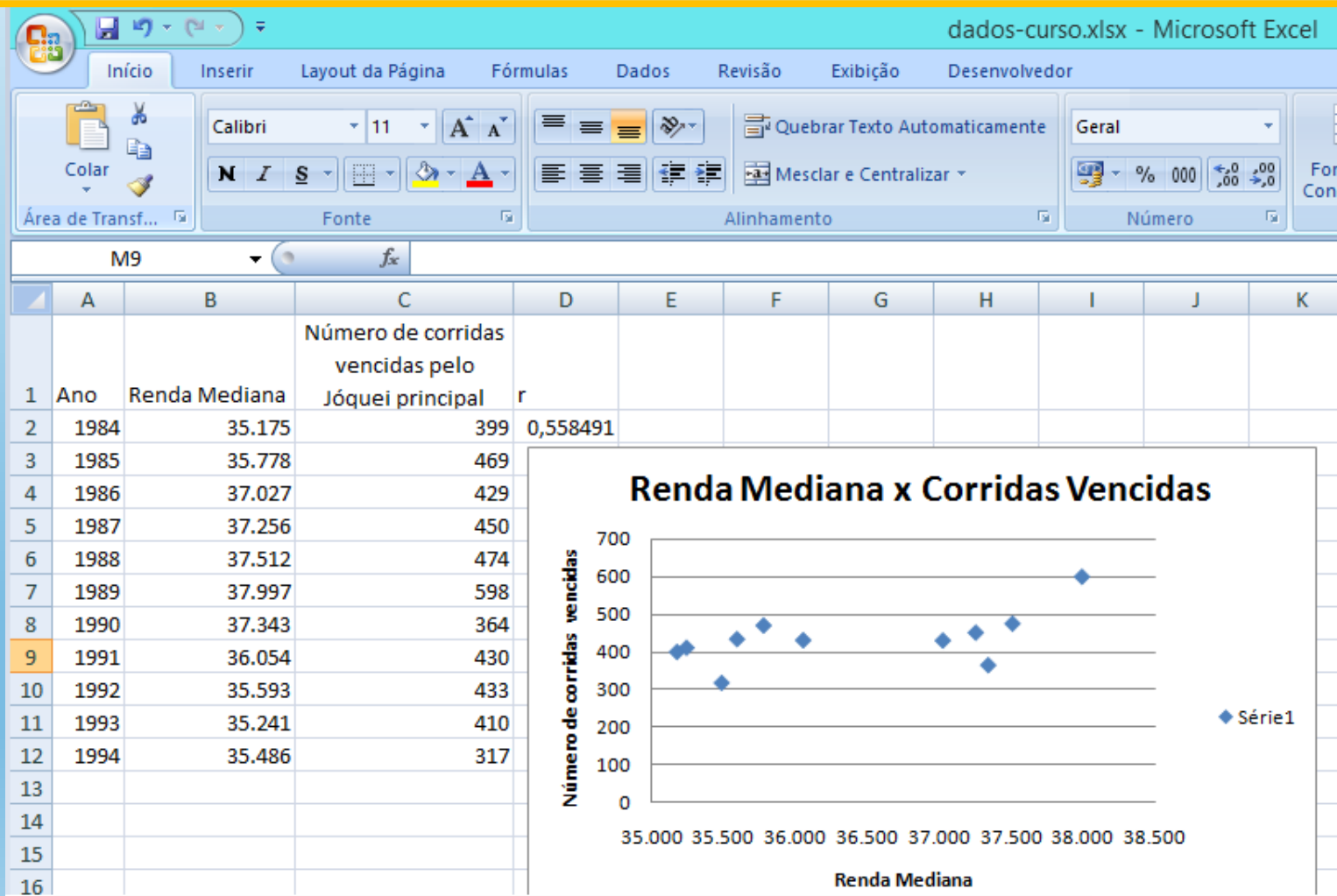
A correlação linear calculada para o exemplo anterior também pode ser expressa através de um gráfico de dispersão. Para gerá-lo, clique na Barra de ferramentas – Inserir – Dispersão



The screenshot shows the Microsoft Excel interface. The 'Inserir' (Insert) ribbon is active, and the 'Gráficos' (Charts) group is expanded. The 'Dispersão' (Scatter) chart type is highlighted with a red circle. A yellow arrow points to the 'Dispersão' icon in the 'Gráficos' group. Below the ribbon, a data table is visible with columns for 'Ano', 'Renda Mediana', and 'Número de corridas vencidas pelo Jôquei principal'.

	A	B	C	D	E
	Ano	Renda Mediana	Número de corridas vencidas pelo Jôquei principal		
1					
2	1984	35.175	399	0,558491	
3	1985	35.778	469		
4	1986	37.027	429		
5	1987	37.256	450		
6	1988	37.512	474		
7	1989	37.997	598		
8	1990	37.343	364		
9	1991	36.054	430		
10	1992	35.593	433		
11	1993	35.241	410		
12	1994	35.486	317		
13					

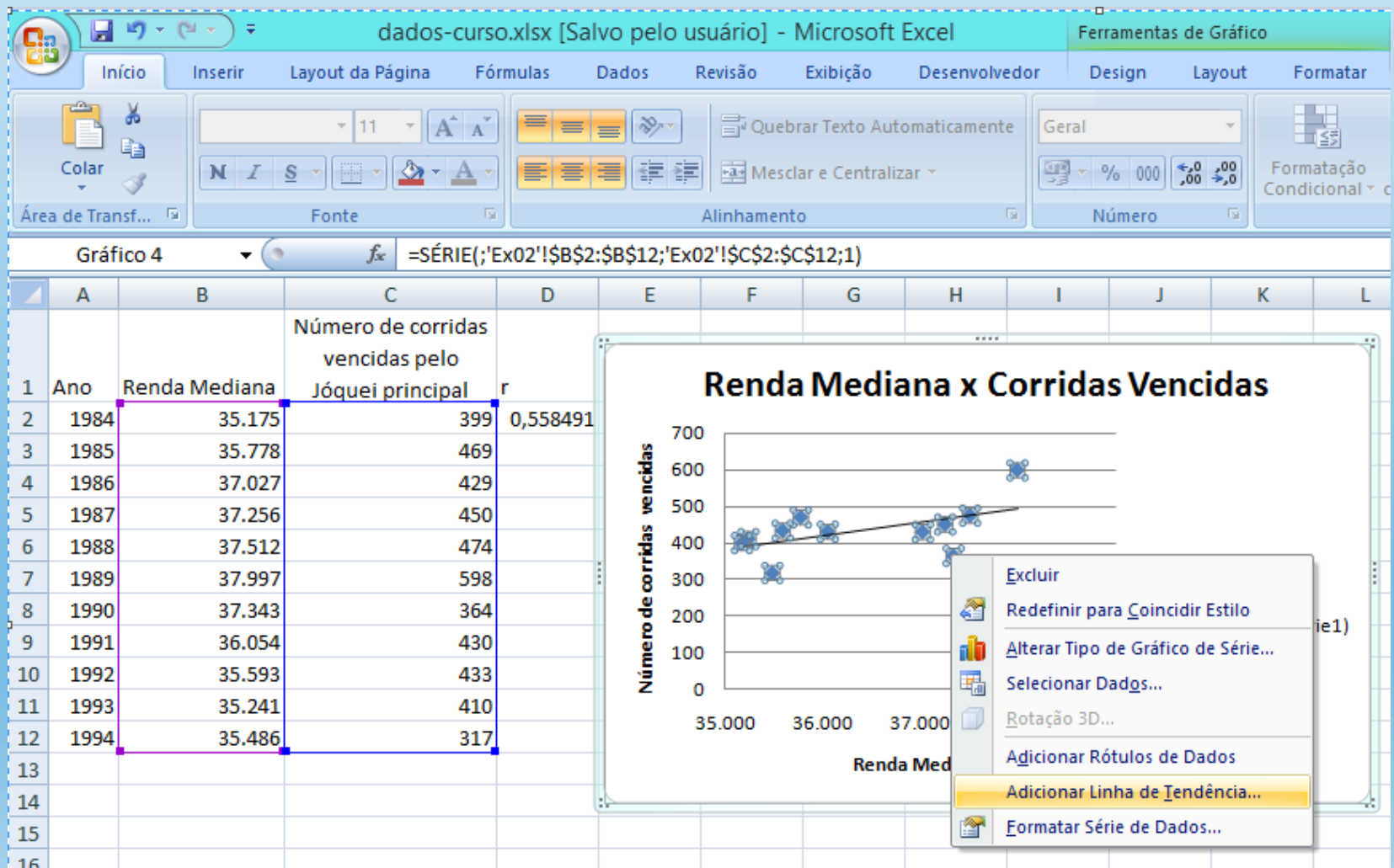
O gráfico de dispersão é bastante útil para demonstrar a existência ou não de relações entre duas variáveis. Quanto mais alinhados estiverem os pontos à reta, maior deve ser a correlação linear entre as duas variáveis. No exemplo utilizado, as duas séries aleatórias mostram o seguinte padrão:



É possível, no mesmo gráfico de dispersão, inserir a reta de regressão de uma variável em relação à outra

1) Clique sobre um dos pontos azuis do gráfico

2) Com o botão direito selecione “Adicionar linha de tendência”



3) Escolher o tipo de ajuste, p. ex., linear

4) É possível exibir a equação da reta linear e o valor de  $R^2$

The image shows a screenshot of Microsoft Excel with a data table and the 'Formatar Linha de Tendência' (Format Trendline) dialog box open. The data table is as follows:

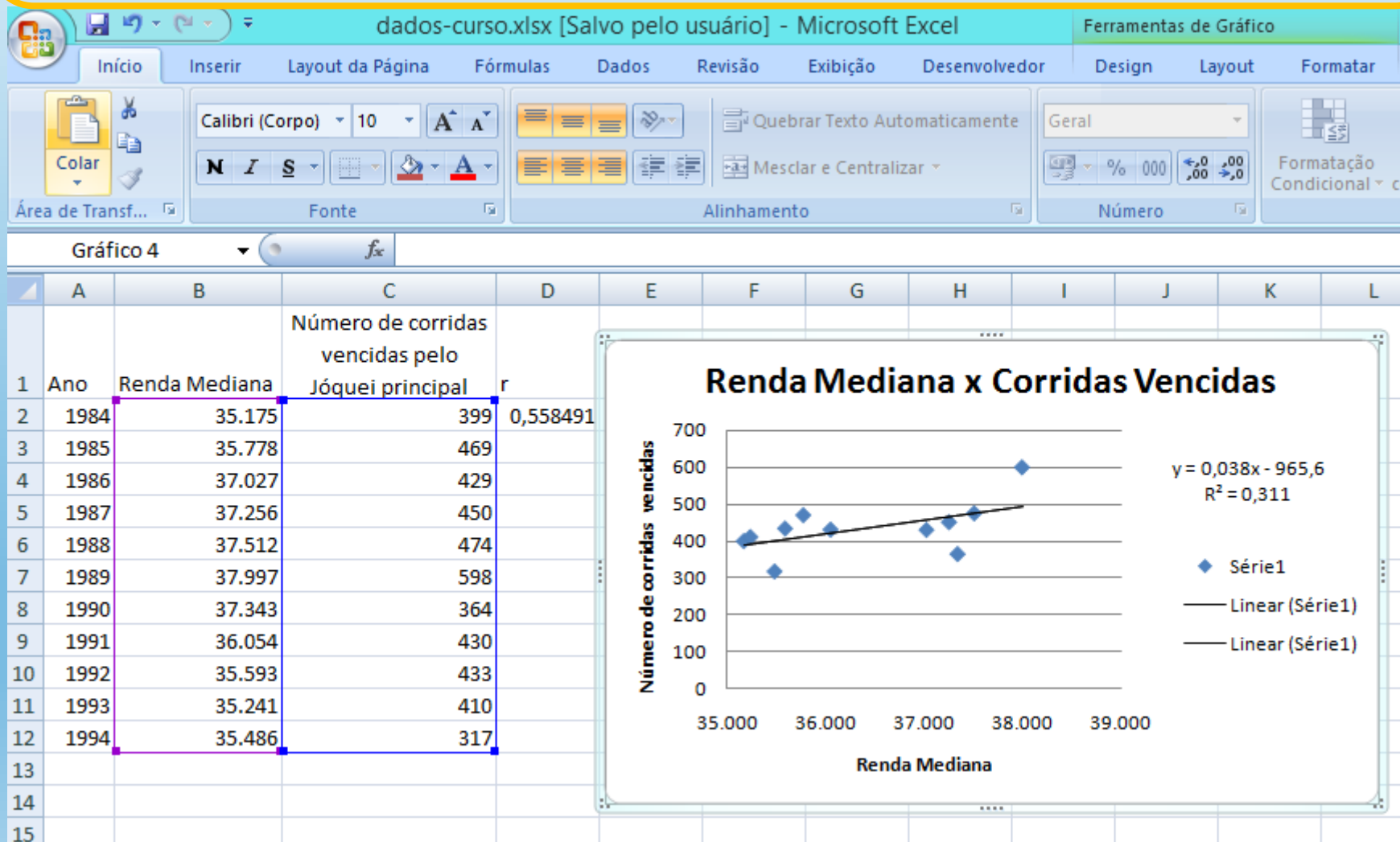
	A	B	C
	Ano	Renda Mediana	Número de corridas vencidas pelo Jôquei principal
1			
2	1984	35.175	399
3	1985	35.778	469
4	1986	37.027	429
5	1987	37.256	450
6	1988	37.512	474
7	1989	37.997	598
8	1990	37.343	364
9	1991	36.054	430
10	1992	35.593	433
11	1993	35.241	410
12	1994	35.486	317
13			
14			
15			
16			
17			
18			
19			
20			
21			
22			

The 'Formatar Linha de Tendência' dialog box is open, showing the following options:

- Opções de Linha de Tendência**
  - Cor da Linha
  - Estilo da Linha
  - Sombra
- Tipo de Tendência/Regressão**
  - Exponencial
  - Linear
  - Logarítmica
  - Polinomial (Ordem: 2)
  - Potência
  - Média Móvel (Período: 2)
- Nome da Linha de Tendência**
  - Automático: Linear (Série 1)
  - Personalizado: [ ]
- Previsão**
  - Avançar: 0,0 períodos
  - Recuar: 0,0 períodos
  - Definir Interseção = 0,0
  - Exibir Equação no gráfico
  - Exibir valor de R-quadrado no gráfico

Buttons: Fechar

Ao terminar de selecionar as opções de formato, clique em fechar  
Os resultados serão exibidos como o modelo abaixo



# COEFICIENTE DE DETERMINAÇÃO $R^2$

- Indica o grau do ajuste linear entre duas variáveis
- Indica o grau de dependência linear entre duas variáveis
- Se uma variável pode ser considerada como preditora em relação a outra

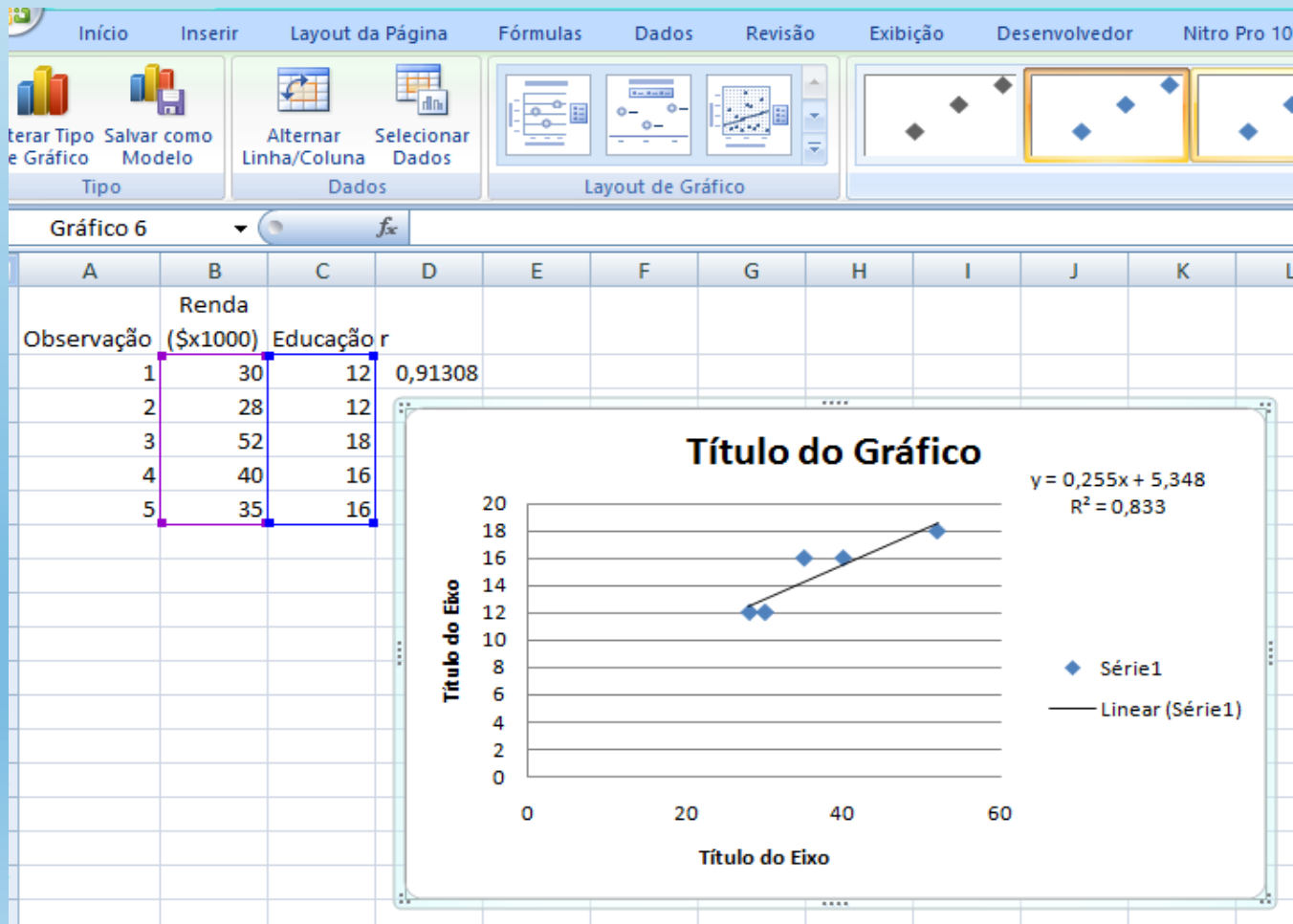
O que é uma variável preditora?



### EXERCÍCIO 03: Seguir os mesmos passos do exercício anterior

- 1) Escolha o formato do gráfico
- 2) Escreva o nome do gráfico
- 3) Coloque nome nos eixos X e Y

O Resultado final será o seguinte:



## **EXERCÍCIO 04**

Utilizem os dados da planilha Ex04 e calculem:

- 1) A correlação entre a série de precipitação e a de OLR**
- 2) Gráfico de dispersão para as variáveis precipitação e OLR**
- 3) Correlação linear entre a precipitação e a TSM**
- 4) Gráfico de dispersão para as variáveis precipitação e TSM**
- 5) Interprete dos gráficos obtidos**

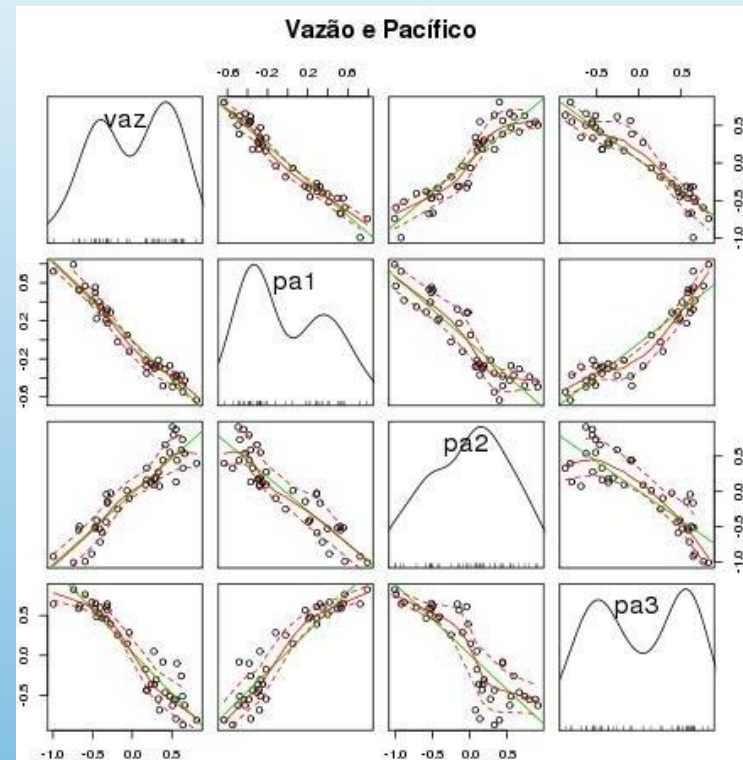
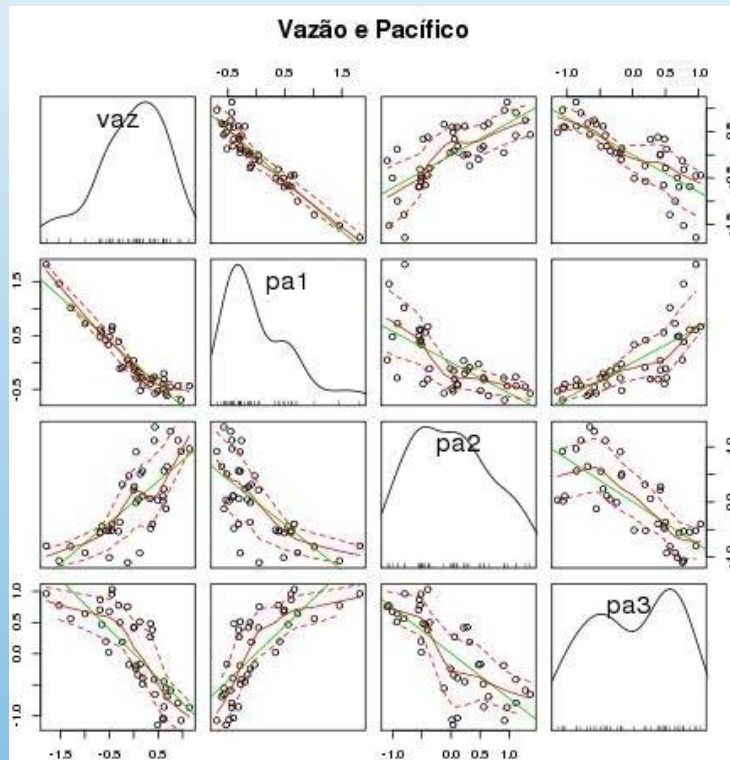
# USO DE OUTROS SOFTWARES ESTATÍSTICOS

## CORRELAÇÃO LINEAR

Outros softwares estatísticos, e gratuitos, tais como o **R**, **GrADS** e **NCL**, são capazes de tratar séries temporais, mas também dados distribuídos espacialmente.

Trazem uma série de recursos gráficos que facilitam a visualização e a geração de saídas mais elaboradas.

# DIAGRAMAS DE DISPERSÃO NO R

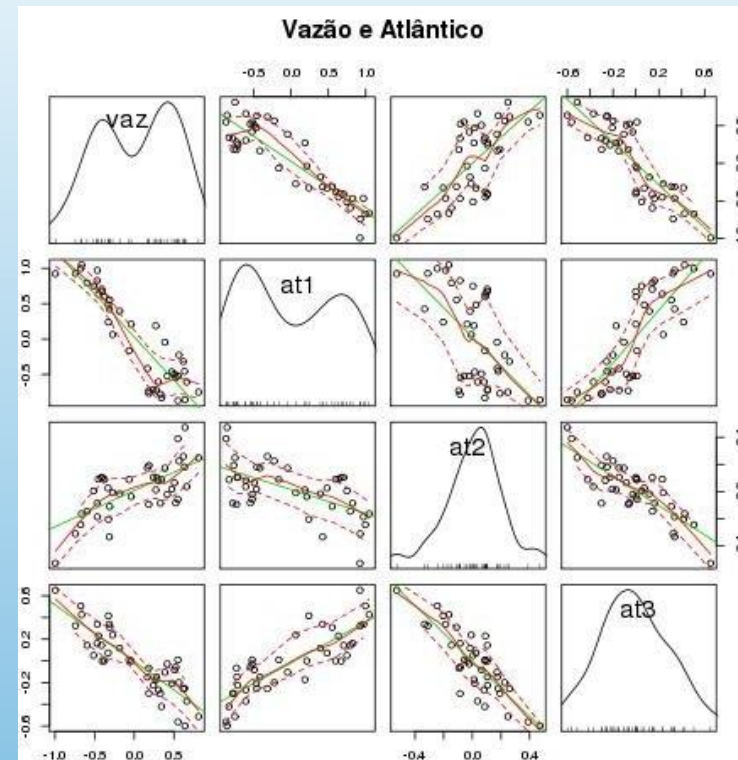
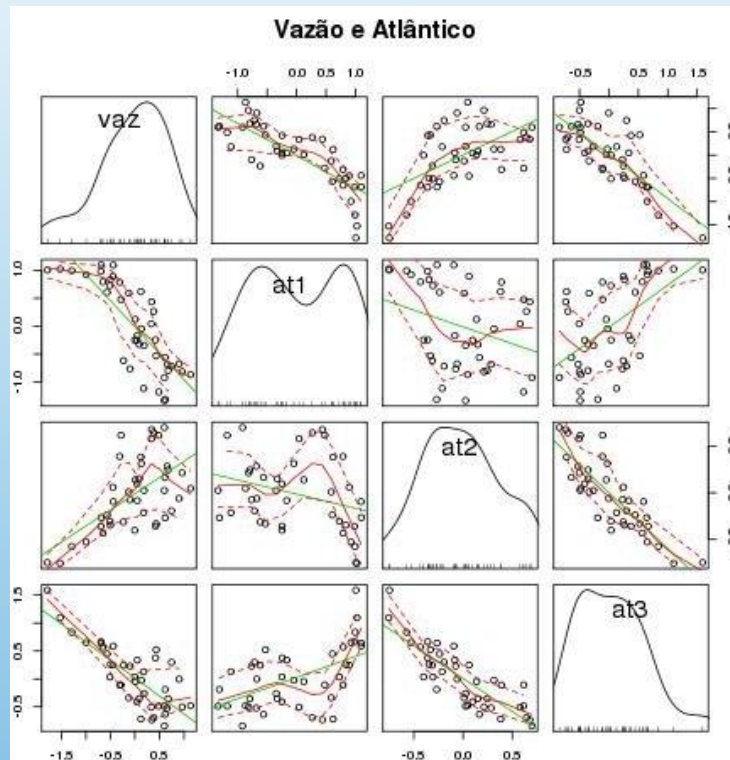


Diagramas de dispersão entre a vazão anual do rio Madeira e a TSM média nas áreas PA1, PA2 e PA3, suavizadas com média móvel (a) 6 e (b) 12 anos.

PA1 PA2 PA3 – áreas oceânicas no Pacífico

Fonte: SILVA, E.R.L.D.G. **Associação da variabilidade climática dos oceanos com a vazão de rios da Região Norte do Brasil**. Dissertação de Mestrado. São Paulo: Universidade de São Paulo. Faculdade de Filosofia, Letras e Ciências Humanas. Departamento de Geografia, 2013. 182p.

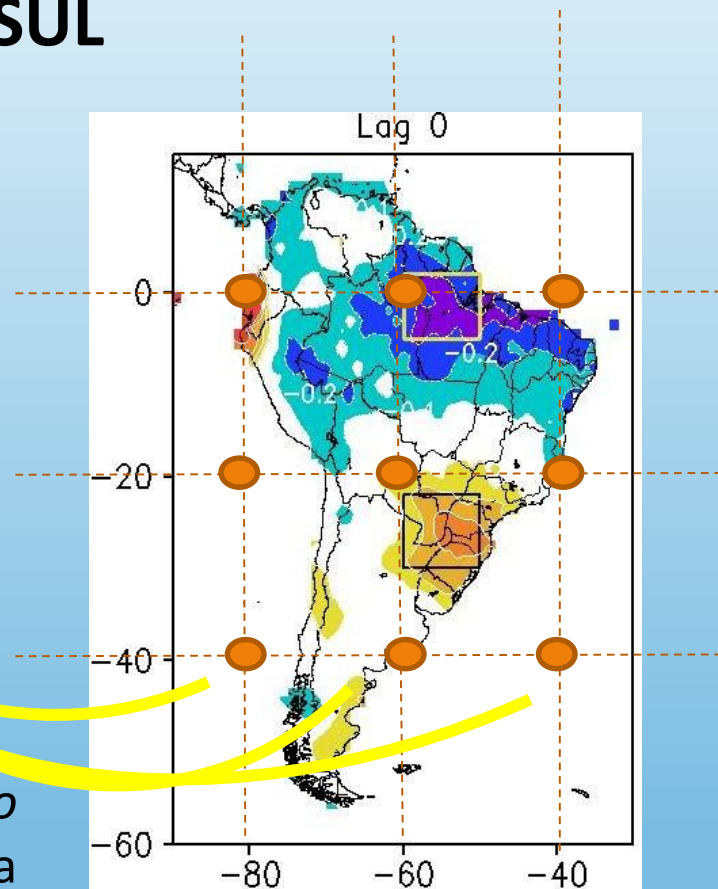
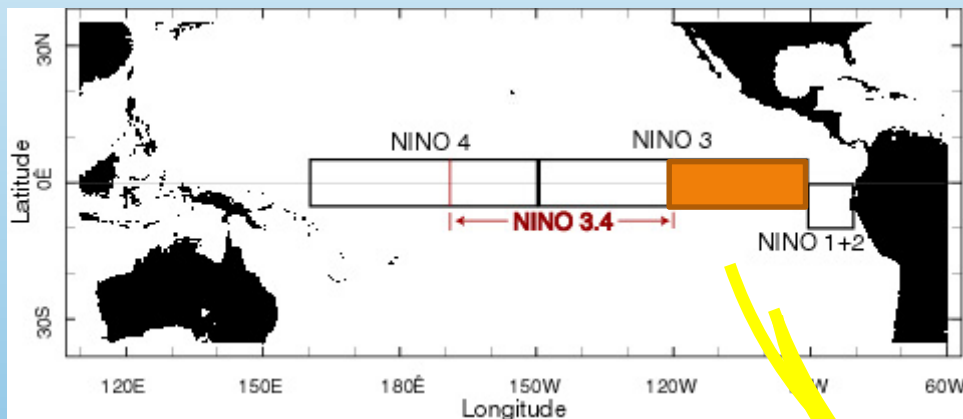
# DIAGRAMAS DE DISPERSÃO NO R



Diagramas de dispersão entre a vazão anual do rio Madeira e a TSM média nas áreas AT1, AT2 e AT3, suavizadas com média móvel (a) 6 e (b) 12 anos AT1 AT2 AT3 áreas oceânicas no Atlântico.

Fonte: SILVA, E.R.L.D.G. **Associação da variabilidade climática dos oceanos com a vazão de rios da Região Norte do Brasil**. Dissertação de Mestrado. São Paulo: Universidade de São Paulo. Faculdade de Filosofia, Letras e Ciências Humanas. Departamento de Geografia, 2013. 182p.

# CORRELAÇÃO LINEAR ESPACIAL TSM DA REGIÃO DE NIÑO 1+2 x PRECIPITAÇÃO NA AMÉRICA DO SUL

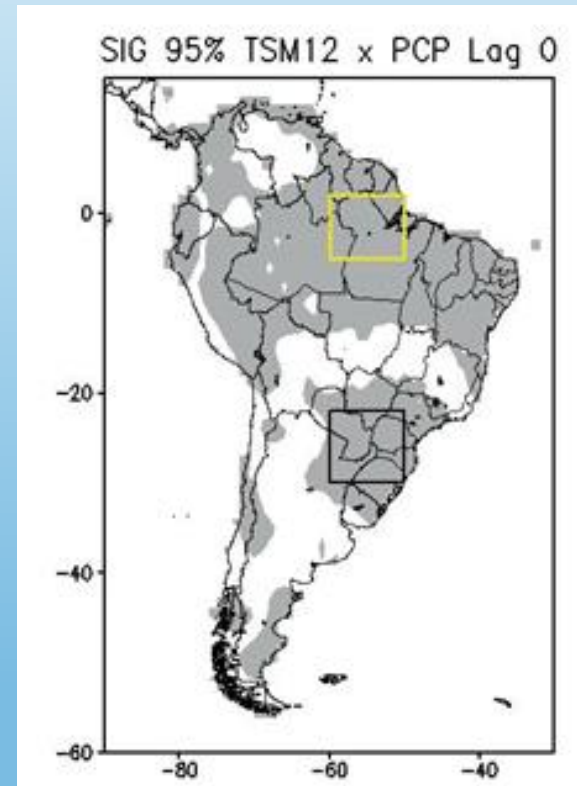


Os valores mensais de TSM das regiões de *Niño* foram correlacionados com os valores da precipitação na América do Sul

# SIGNIFICÂNCIA ESTATÍSTICA

A significância estatística do cálculo do coeficiente de correlação foi avaliada com a aplicação do teste *t-Student*, cujo valor limite para se considerar o cálculo significativo é definido, segundo Costa Neto (1977), por:

$$t_{n-2} = r \frac{\sqrt{n-2}}{\sqrt{1-r^2}}$$



# SIGNIFICÂNCIA ESTATÍSTICA

É um valor que expressa a confiabilidade estatística referente a um cálculo estatístico

{ média  
correlação  
tendência linear

**Como definimos se  $r = 0,6$  é um valor estatisticamente confiável de correlação linear para os dados usados?**

**Resp.: Dependerá do valor de  $r$  e de  $N$ , como indicado na equação anterior.**



# SIGNIFICÂNCIA ESTATÍSTICA

- Esta pergunta deve ser feita para fornecer alguma garantia relativa ao valor obtido para determinada estatística, que indique que o valor resultante não advém da aleatoriedade.
- Esta garantia pode ser expressa através de níveis de confiança:

90%, 95%, 99%

são níveis de confiança usados corriqueiramente.

# SIGNIFICÂNCIA ESTATÍSTICA

- Existem alguns testes de significância mais usados: teste t-Student (supõe a distribuição normal dos dados)
- Para tanto, precisamos saber qual é a quantidade de valores usados no cálculo da estatística (n) e qual é o valor obtido da estatística (r, no caso do coeficiente de correlação)

para coeficiente de correlação

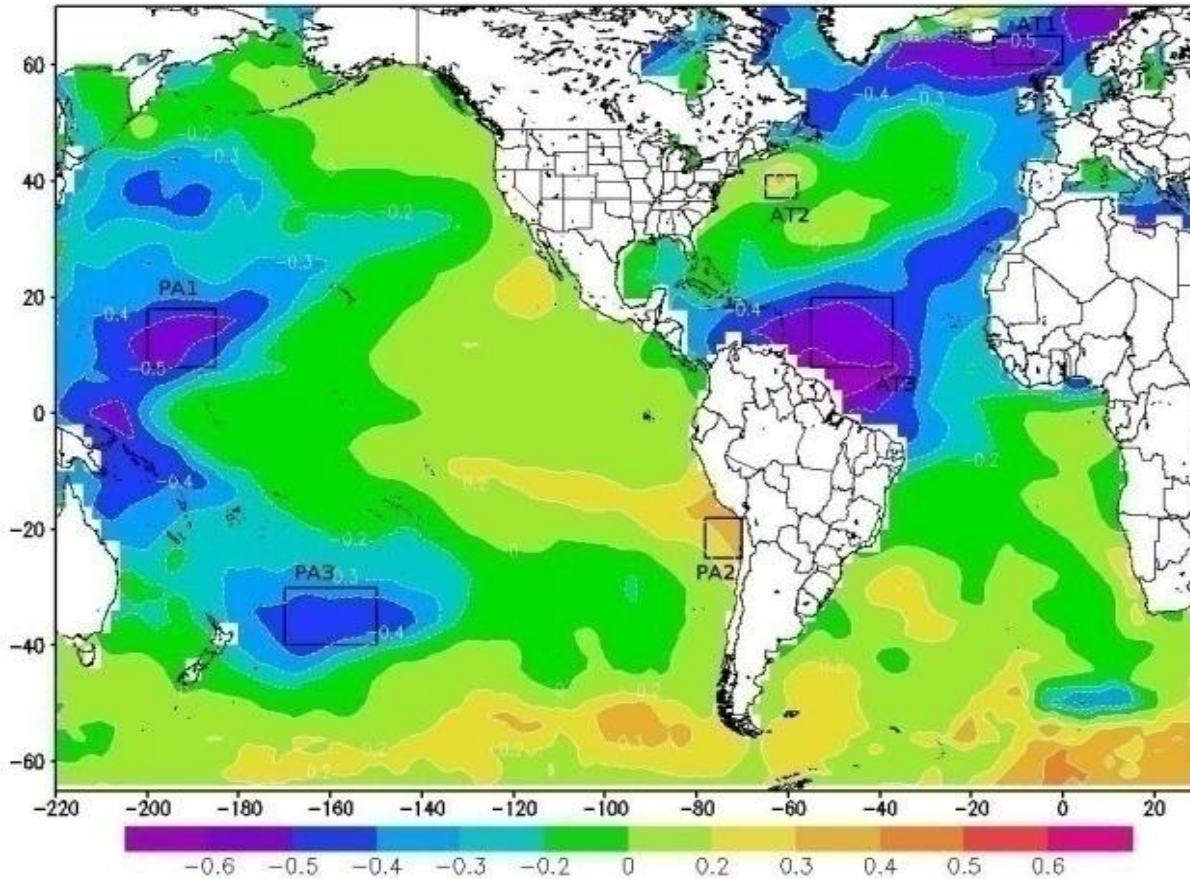
$$t_c = t_{n-2} = r \frac{\sqrt{n-2}}{\sqrt{1-r^2}}$$

$t > t_c \rightarrow$  cálculo estatisticamente significativo

$t < t_c \rightarrow$  cálculo não é estatisticamente signif.

# CORRELAÇÃO LINEAR ESPACIAL TSM GLOBAL x VAZÃO DO RIO MADEIRA

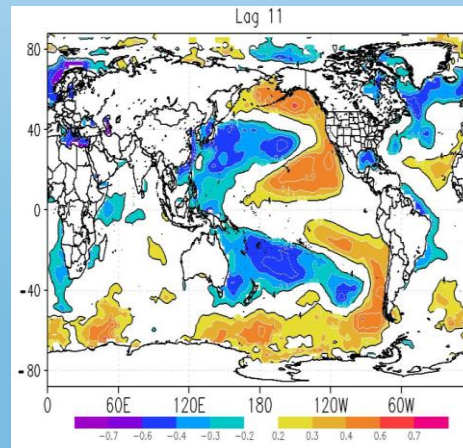
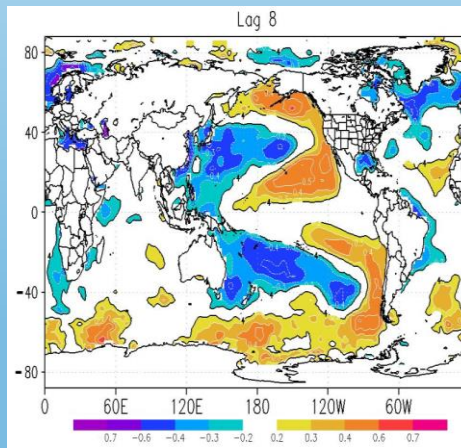
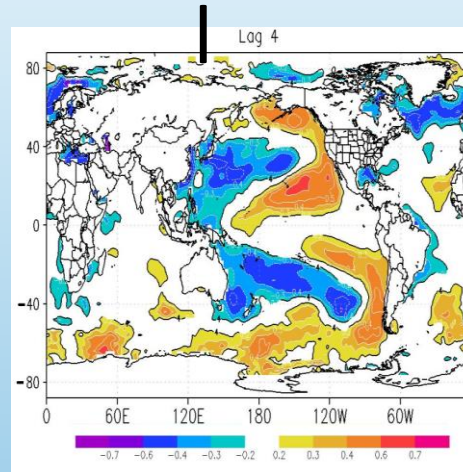
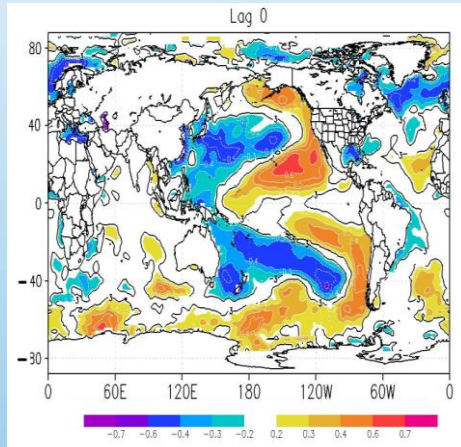
**EXEMPLO 06** VAZAO MADEIRA Jan Lag 0



Qual a interpretação que pode ser feita do mapa ao lado?

Fonte: SILVA, E.R.L.D.G. **Associação da variabilidade climática dos oceanos com a vazão de rios da Região Norte do Brasil**. Dissertação de Mestrado. São Paulo: Universidade de São Paulo. Faculdade de Filosofia, Letras e Ciências Humanas. Departamento de Geografia, 2013. 182p.

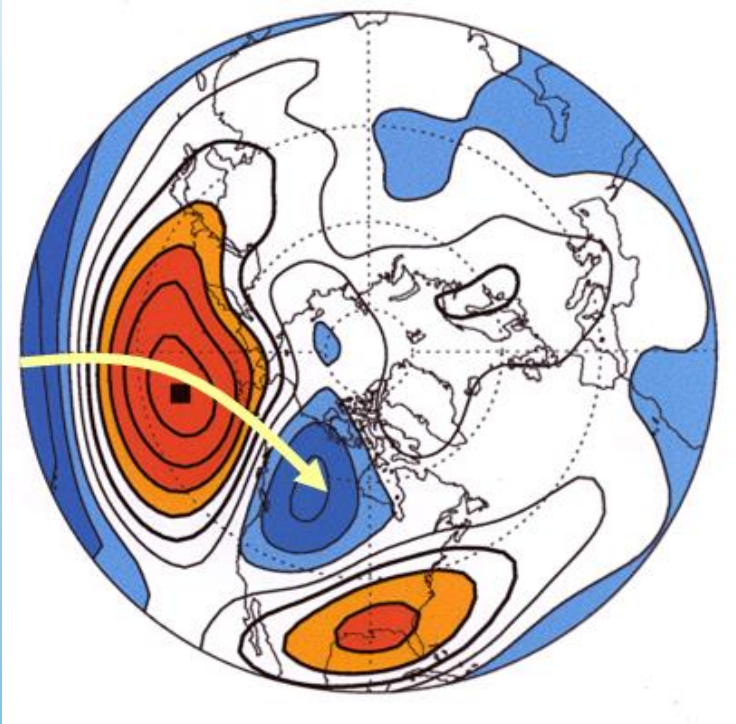
# CORRELAÇÃO LINEAR ESPACIAL TSM GLOBAL x VAZÃO NO PANTANAL



Qual a interpretação  
que pode ser feita  
dos mapas ao lado?

*Lagged linear correlation between Pantanal discharge and SST monthly data for the period 1970-2003, for (a) lag=0, (b) lag=4 (c) lag=8 and (d) lag=11 months. The first month in SST time series is always January. The statistical significant areas at 99% (t-Student test) are given by the black lines. (Silva et al., 2016 TAAC)*

# ALTURA GEOPOTENCIAL 500 mb



Qual o padrão que pode ser observado através da correlação da altura geopotencial em 500 mb com o valor no Pacífico Norte?  
(Resp.: PNA)

Spatial distribution of correlation of the 500 mb geopotential height anomaly time series (Seasonal JFM) at all points on the Northern hemisphere with the time series at a specified “base point” - North Pacific. Red colors positive correlation, blue colors negative correlation. Yellow arrow indicate meridional orientation of spatial structure existing in the correlation pattern. Picture courtesy of Prashant Sardeshmukh, CDC/OAR

# Script GrADS - Correlação Linear

script no grads – arquivo texto com qualquer nome

```
'c'  
'reinit'  
'set display color white'           ! Define fundo branco para a figura  
'c'                                  ! clear  
'set grads off'  
'sdfopen cru_ts3.20.1901.2011.pre.dat.nc' ! abre arquivo nc  
'set y 1'                            ! fixa uma latitude  
'set z 1'                            ! fixa um nível atmosférico  
'set t 601 1332'                    ! fixa o período de tempo  
'define AS = aave(pre, lon=-90, lon=-30, lat=-60, lat=20)' ! calcula a média de pre em um retângulo  
'set lon -90 -30'                   ! define domínio lon  
'set lat -60 20'                    ! define domínio lat  
'set z 1'                            ! define nível atmosférico  
'set t 601'                          ! fixa um tempo  
'set gxout shaded'                   ! define forma mapa  
'set clevs 0 5 7.5 10 12.5 15 17.5 20 22.5' ! define níveis da variável  
'set ccols 49 47 45 42 41 23 24 25 27 29' ! define níveis de cores
```

## .... continuação do script

```
'set rgb 49 20 100 210'  
'set rgb 47 40 130 240'  
'set rgb 45 80 165 245'  
'set rgb 42 180 240 250'  
'set rgb 41 225 255 255'  
'set rgb 23 255 192 60'  
'set rgb 24 255 160 0'  
'set rgb 25 255 96 0'  
'set rgb 27 225 20 0'  
'set rgb 29 165 0 0'  
  
'd tregr(AS, pre, t=601, t=1332)*10'  
'define coeff = tregr(AS, pre, t=601, t=1332)'  
'define preave = ave(AS, t=601, t=1332)'  
'define ASave = ave(AS, t=601, t=1332)'  
'd (coeff * (AS - ASave) + preave)/10'  
'set gxout bar'  
'cbarn'  
'set strsiz .20'  
'set string 1 c 5 0'  
'draw string 5.5 8 COEFICIENTE TREGR 1951-2011'  
'printim tregr-shaded.png'
```

# EXERCÍCIO GrADS

## Correlação precipitação na América do Sul

1) Calcule a correlação linear entre os índices climáticos ODP (Oscilação Decadal do Pacífico), MEI e IOS (Índice da Oscilação Sul) e a precipitação mensal na América do Sul. Analise os resultados.

- a) Baixe os índices climáticos de ODP, IOS e MEI do site do CDC-NOAA;
- b) Descreva o significado de cada índice;
- c) Faça a correlação linear entre os índices climáticos e a precipitação mensal na América do Sul;
- d) Analise os 3 mapas resultantes.